Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Sentiment Analysis and Topic Modeling on News Headlines

Vijay Yadav[1], Subarna Shakya[2]

[1,2]Department of Electronic and Computer Engineering, Institute of Engineering, Tribhuvan University, Lalitpur, Nepal

E-mail: [1]076mscsk020.vijay@pcampus.edu.np, [2]drss@ioe.edu.np

Abstract

Sentiment analysis and topic modeling has wide range of applications from medical to entertainment industry, corporates, politics and so on. News media play vital role in shaping the views of public towards any product or people. The dataset used for this work is news headlines dataset of one of the leading new portals of India i.e., Times of India. This research aims to perform comparative study of both supervised and unsupervised learning for text analysis and use the best performing models in both the category for prediction of sentiment and topic classification of news headlines. For sentiment analysis, supervised techniques like Machine learning ensemble model and Bi-LSTM have used. Similarly, unsupervised techniques like LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Analysis) have been for topic modeling.

Keywords: Sentiment analysis, Topic modeling, Data visualization, Bi-LSTM, LDA, LSA

1. Introduction

Sentiment analysis is considered as one of the important aspects in many areas like treatment of mental health problem, understanding the sentiment of people regarding the product of any companies, prediction of share markets, views of people regarding the launch of any scheme by the government, detecting fake news being circulated among the people, and so on. There are various ways in which one can try to understand the sentiment of people like conducting survey, preparing questionnaire, using the audio or video of the individuals or using the posts made by them on various social media sites.

Sentiment analysis is an open research field. Various researches have been done and many research works are still going on regarding sentiment analysis. All researches offer something different than other but none of them are have been able to offer a complete solution

DOI: https://doi.org/10.36548/jucct.2022.3.008
Received: 10.08.2022, received in revised form: 26.08.2022, accepted: 12.09.2022, published: 22.09.2022

and this is also due to the limitations of research in the field of Natural Language Processing and unavailability of open and good datasets in the regard of this field. Sentiment analysis can be done through datasets available in various forms such as text, audio, video, images and so on. Comparatively it is easier to perform analysis on textual data due to the limitations of hardware and other resources for training a model. Sentiment analysis is an important field of NLP. NLP plays vital role in understanding the context of textual data.

Topic modeling is another technique for text analysis. It is an unsupervised technique to analyse the given set of documents. As the name suggests, it is used to discover number of topics within the given sets of text like tweets, books, articles and so on. Each topic consists of consists of words where order of the words does not matter. It performs automatic clustering of words which best describes a set of documents. It gives us insight into number of issues or features users are taking about as a group.

For example, let us say a company has launched a software in the market and it receives a number feedbacks regarding various features of the product within a specified time period. Now, rather than going through each review one by one, if we apply topic modeling, we will come to know how users have perceived the various features of the product very quickly. It is one of the important techniques to perform text analysis on unstructured data. After performing topic modeling, we can even perform topic classification to predict under which topic the upcoming reviews fall. There are various techniques to perform topic modeling, among which LDA is considered to be the most effective one.

2. Related Work

Though the social media platforms have provided freedom of expression to people, it has led to cause some disturbing events as well. The hate speech over social media is one of the major concerns for various users. It has caused serious threat to the life of people. Sentiment analysis can play a big role in identifying hate speech over social media platforms and punch the perpetrators of those hate speeches. For this various deep learning techniques like CNN, RNN can be used [1].

Due to this openness and freedom provided by social media, it becomes the duty of those platforms to do their bits in helping deal with existing and increasing mental health problems. Still, none of the researches have assured that it can help to accurately predict all the suicidal and non-suicidal posts made through social media by different users [2]. It is due to

the problems associated with Natural Language Processing. Except social media posts, there are also some other ways that could contribute to sentiment analysis of individuals. One of them could be classification of suicidal notes. It involves linguistic analysis. Different people express their feeling differently in the suicide notes.

The use of language, pronouns, grammars vary vastly which makes it difficult to correctly classify suicidal and non-suicidal notes [3]. Some other researches have also been conducted to identify genuine and fake suicide notes. From traditional to latest machine learning techniques, a lot have been used in this problem domain. Use of NLP and text mining in sentiment analysis is still in its infancy, that's why wecould see that the big social media tech giants find it difficult to accurately identify posts expressing negative sentiments [4].

In the early days of sentiment analysis, classifications were done based on the individual words rather than understanding the context in which those words have occurred [5]. Later on, wordnet based approach was proposed for sentiment analysis by calculating the distance between the word appeared in a text to the word "good" or "bad" [6]. Some researchers even used Cosine distance for better accuracy. The use of Bi-LSTM in sentiment analysis takes the context into consideration due to which it has been found to be more effective in sentiment analysis over other methods [7].

It is not easy to deal with huge text and get insights into what the text is trying to depict, topic modeling is one of the techniques to easily under the subjects depicted by the large collection of text [8]. Topic modeling has been studied and applied in various fields like political science, customer reviews, software engineering, medical, linguistic science [9]. Though there are many topic modeling algorithms available, it is necessary to perform the tuning and optimization of such algorithms to get reliable result. It is necessary to understand the underlying process in topic modeling algorithms to decide which algorithm best suit the given purpose [10].

3. Proposed Work

The dataset used for this research work is "times_of_india_news_headlines" as shown in figure 2. The dataset has been taken from Kaggle. This dataset contains a number of collection of headlines of the year 2019 AD. The dataset mainly covers the news of events in India. It has three columns and over 60 thousand of rows. As the dataset is not cleansed, it

requires preprocessing task to be performed before applying it on model. The columns with the datasets are Text (i.e., news headlines)**,** Published date and Sentiment**.**

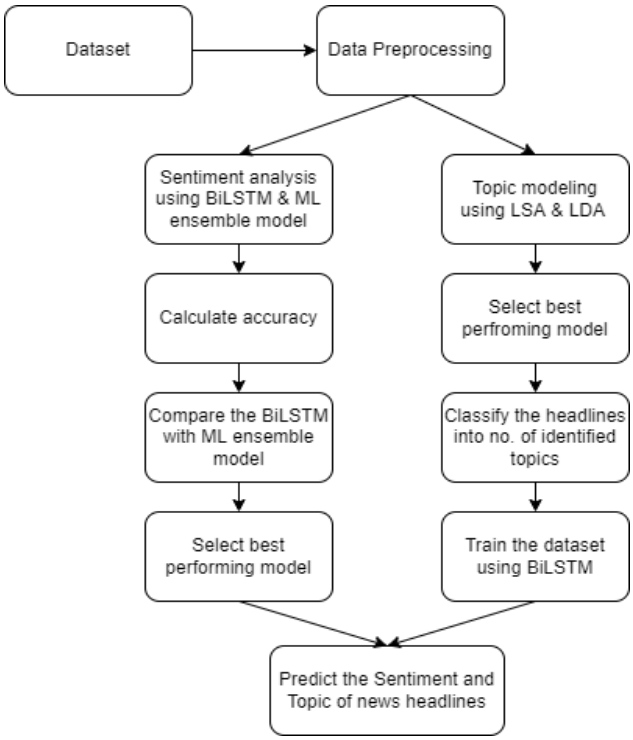

Figure 1. Methodology

3.1 Dataset used

text	published_date	Sentiment
MP to study new Motor Act before implementing ...	2019-09-03	Negative
Bhujbal says higher power rate driving industr...	2019-12-10	Negative
Mandira Bedi shares a monochrome picture in a ...	2019-12-12	Negative
Harassment in office: DM seeks action-taken re...	2019-11-18	Negative
Two students end their lives	2019-06-09	Negative
...	...	...
Going to India Gate tonight? You could be in f...	2019-12-31	Positive
40;000 cops to ensure safe New Year celebratio...	2019-12-31	Positive
Ten more private trains from Mumbai on cards	2019-12-31	Positive
A trip to a tribal village with super cop Vija...	2019-12-31	Positive
Wi-Fi at over 5;500 railway stations; thanks t...	2019-12-31	Positive

Figure 2. Sample of dataset

3.2 Sentiment analysis

3.2.1 RNN

In traditional neural networks, inputs and outputs were independent of each other. To predict next word in a sentence, it is difficult for such model to give correct output, as previous words are required to be remembered to predict the next word. RNN remembers everything. It overcomes the shortcomings of traditional neural network with the help of hidden layers.

Because of the quality to remember the previous inputs, it useful in prediction of time series. This is called Long Short-Term Memory (LSTM).

Combining two independent RNN together forms a Bi-LSTM (Bidirectional Long Short-Term Memory) as shown in figure 3. It allows the network to have both forward and backward information. Bi-LSTM gives better result as it takes the context into consideration.

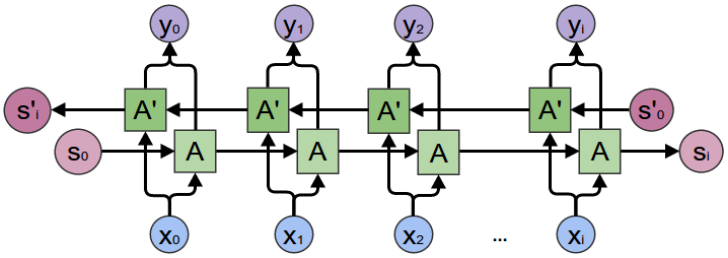

Figure 3. Bi-LSTM Architecture. Source: colah's blog

3.3 Topic Modeling

Topic modeling is an unsupervised method used to perform text analysis. When we are given large sets of unlabelled documents, it is very difficult to get an insight into the discussions upon which the documents are based upon. Here comes the role of topic modeling. It helps to identify a number of hidden topics within a set of documents.

3.3.1 LDA

LDA (Latent Dirichlet Allocation) is one of the most used and well-known technique to perform topic modeling. The word latent means hidden because we don't know what topicsa set of documents contain. Based on probability distribution of occurrence of words in each document, they are allocated to defined topics.

a) Working of LDA

- What we want for LDA is to learn topic mix in each document and also learn the word mix in each document.
- We choose a random number of topics for the given dataset.
- Assign each word in each given document to one of the defined topics, randomly.
- Now, we go through each and every word and to which topic those words are assigned to in each of document. Then, it is analyzed how often the topic occurs in the document and how often the word occurs in the topic as a whole. Base on this analysis, new topic is assigned to the given word.

- It goes through a number of such iterations and finally the topic will start making sense. We can analyze those found topics and assign a suitable name to those topics which best describe them.

3.3.2 LSA

LSA (Latent Semantic Analysis) is another technique used for topic modeling. The main concept behind topic modeling is that the meaning behind any document is based on some latent variables so we use various topic modeling techniques to unravel those hidden variables i.e., topics, so that we can make sense out of given document. LSA is mostly suitable for large sets of documents. It converts the documents into document term matrix before actually deriving topics from the documents.

b) Working of LSA

- The given text is converted into document- term matrix using either bag of words or Term Frequency- Inverse Document Frequency.
- Then, using Truncated Singular Value Decomposition (SVD). It is at this stage the topics within the documents are identified. Mathematically, it can be given as,

$$A = U_t S_t V_t^T$$

In simple terms what the above formula represents is that it simply decomposes high dimensional matrix into smaller matrices i.e., u, s and v, where,

A = n*m document-term matrix (n = no. of documents and m = no. of words)U = n*r document-topic matrix (n = no. of documents and r = no. of topics) S = r*r matrix (r = no. of topics)

V = m*r word-topic matrix (m = no. of words and r = no. of topics)

- Finally, we can now classify which document belongs to which topics.

4. Results and Discussion

From the preprocessed result, it was observed that the average no. of words per sentence has reduced to approximately 7 from 12. This helped in feature reduction and speeds up the analysis process.

4.1 Results of sentiment analysis

4.1.1 Positive Vs Negative Frequency Graph

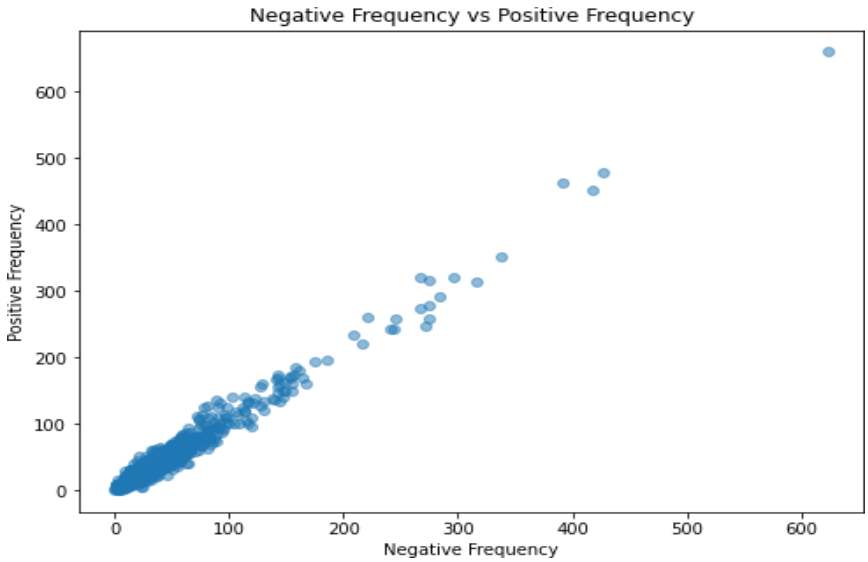

Figure 4. Positive vs Negative frequency

There are some words which have occurred in both positive and negative news headlines. The above figure 4 shows the plot of positive and negative frequency of various words.

4.1.2 Accuracy of sentiment analysis Ensemble model (Using Machine Learning algorithms)

After performing classification using various Machine Learning algorithms, accuracy of those algorithms was used to form an ensemble model using Voting classifier model. The ML models used were Logistic regression, Ridge classifier, Linear SVC, Ada boost, Passive aggressive classifier. The accuracy of ensemble model using both voting hard and voting soft are given below:

Accuracy with Voting hard = 81.67%

Accuracy with Voting soft = 80.45%

The trigram with TF-IDF word embedding was used to train the model. Trigram helps to take the preceding and the succeeding words into consideration. TF-IDF increases the importance of rare words occurring in news headlines. The graph for comparison of accuracy of various n-grams with various word embeddings have been given in figure 5.

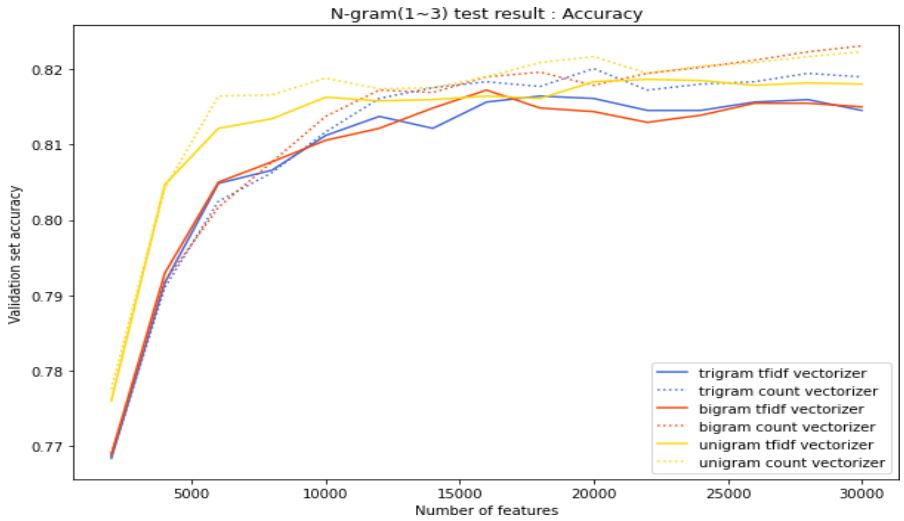

Figure 5. Accuracy comparison for various n-grams

There is not much difference among the accuracies so we take trigram TF-IDF vectorizer for better analysis of the context of the news headlines.

4.1.3 Bi-LSTM model accuracy for sentiment analysis

Precision = 75.38%

Recall or Sensitivity = 84.57%

Model Accuracy = 84.92%

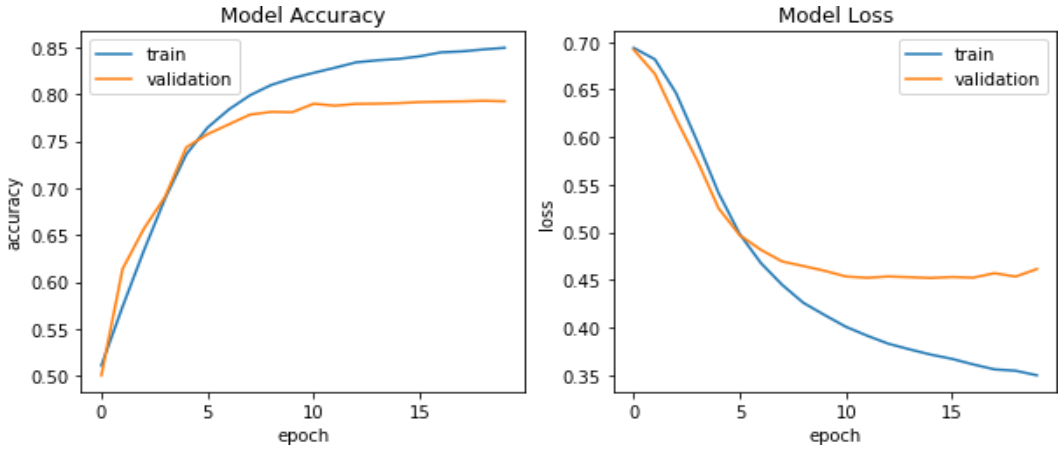

Figure 6. Accuracy and loss graph for Bi-LSTM sentiment analysis model

From the above result for sentiment analysis, we can see the Bi-LSTM model (with accuracy 84.92%) clearly leads the machine learning ensemble model (with accuracy 81.67%) for sentiment analysis. So, Bi-LSTM model was selected for making sentiment prediction of news headlines.

4.2 Results of Topic Modeling

4.2.1 LSA clustering graph for varying number of topics

The figure 7 shows the clustering graph for varying number of topics. The first graph is for 4 topics and keeps on increasing in a multiple of four till 36 topics.

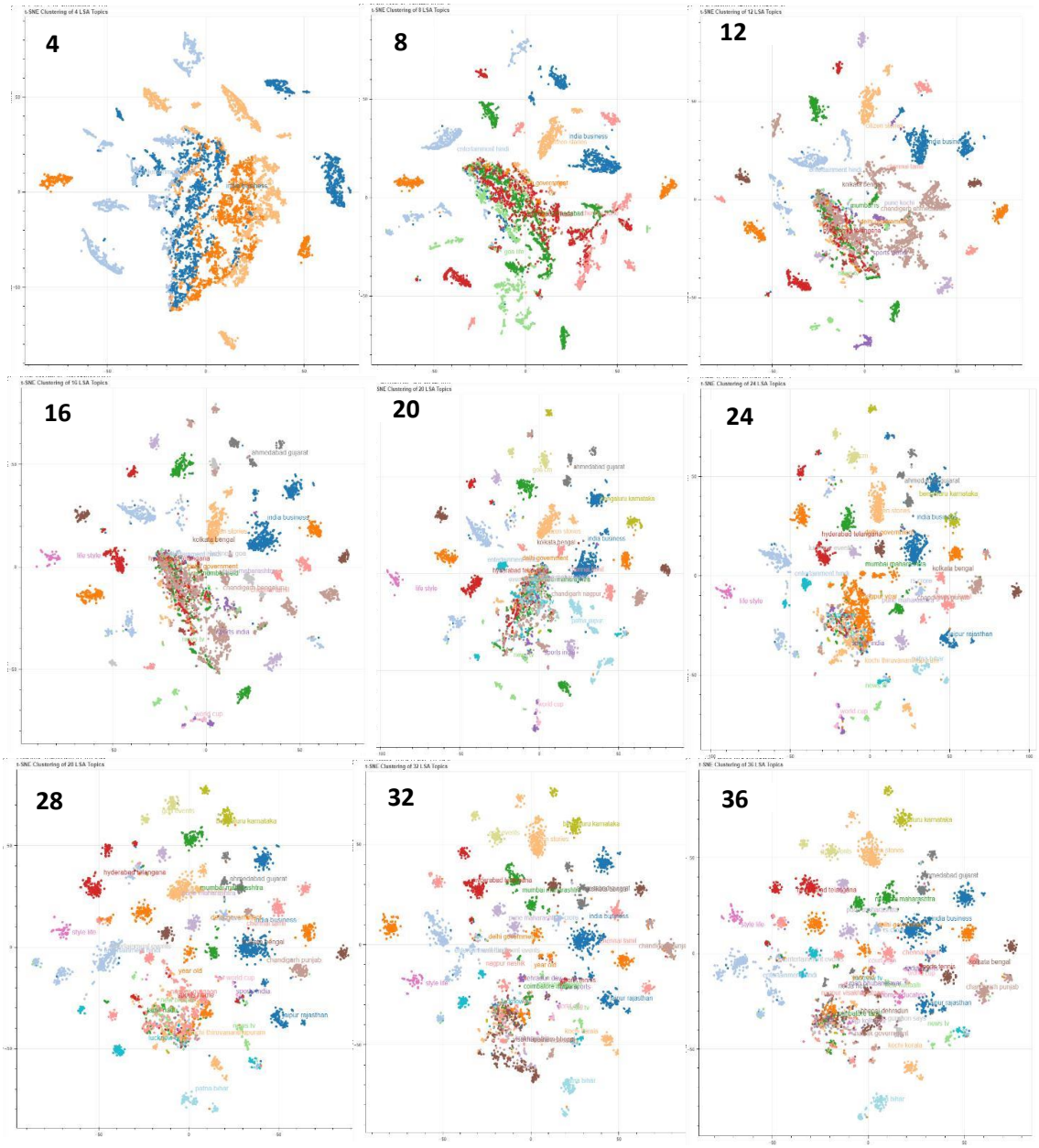

Figure 7. LSA clustering graph for varying number of topics

From the above figure 7, we can observe that the topics become more distinctly separable as the number of topics increases but after a point of time the topics start mixing up with each other, especially after 24 topics in the above figure.

4.2.2 LDA clustering graph for varying number of topics

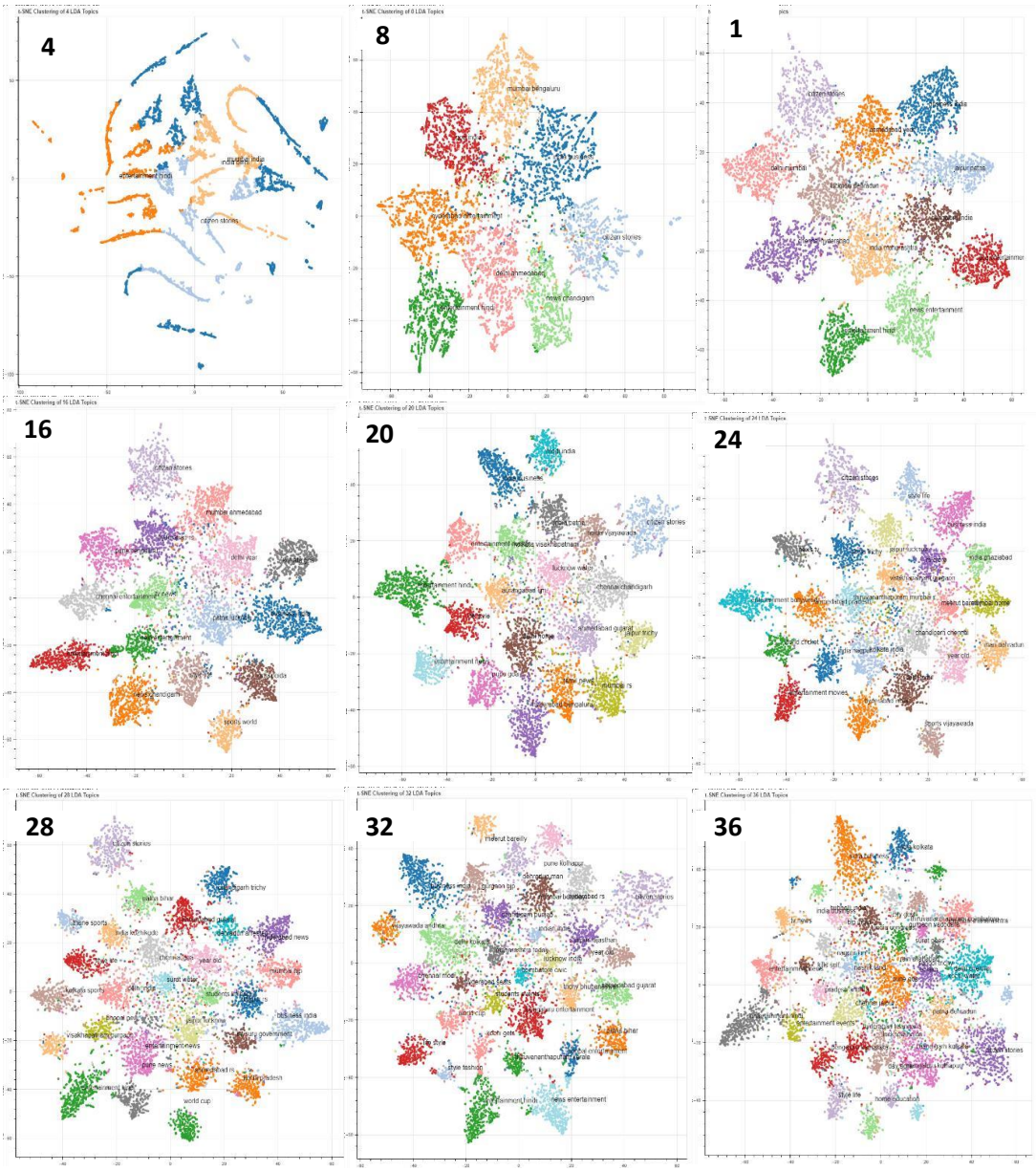

Figure 8. LDA clustering graph for varying number of topics

From figure 8, it can be observed that the topics separation is distinct for certain number of topics but as in the case of LSA, the topics start mixing up with each other especially after 24 topics. But when LDA graph is compared with LSA graph, the topics classification by LDA is better than LSA as there is more clarity and distance among the topics in LDA.

4.2.3 Bar graph for topics distribution

Comparing the topics distribution graph for varying number of topics for both LSA and LDA, it can be observed that LDA has more uniform topic distribution than LSA. So, from

topic distribution graph also, we can conclude that LDA is better algorithm for topic distribution than LSA.

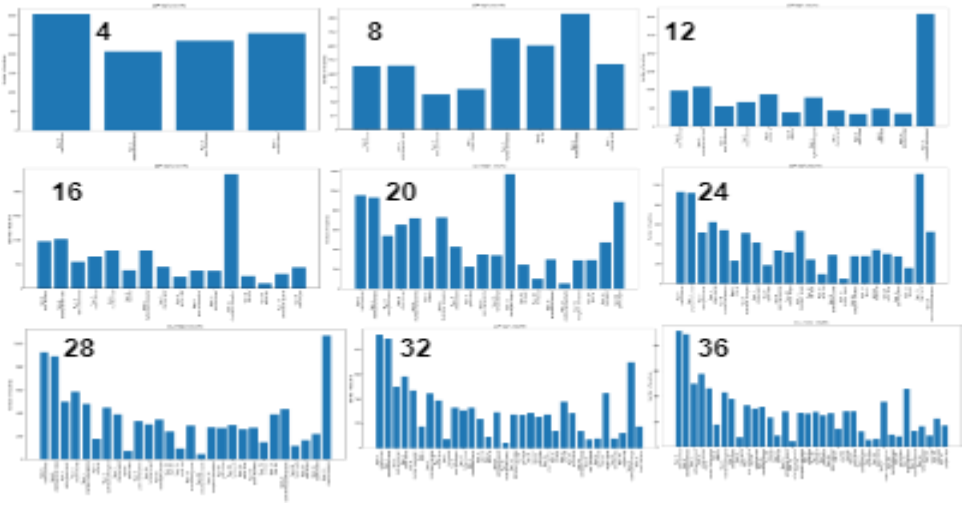

Figure 9. LSA Topics distribution

Figure 10. LDA Topics distribution

From our above analysis, LDA was found to be better for topic modeling so we will use LDA for identifying topics from the given data of news headlines. The entire news headlines would be classified into identified number of topics. Then the Bi-LSTM would be used to train the model in order to make predictions about the topic of particular news headlines.

Though we can have a desired number of topics, but it is necessary to have a balance between number of topics and accuracy of the model. The accuracy of model would be higher for lesser number of topics but as the number of topics are increased, the accuracy of model decreases because the homogeneity of topics distribution also decreases as shown in the figure

10. Thus, we have to make a trade-off between the number of topics and the accuracy of the model. The number of topics should be chosen in such a way that it gives proper insight aboutthe data.

Table 1. Accuracy for varying number of topics

Number of Topics (usingLDA)	**Accuracy % (using Bi-LSTM)**
4	92.94
8	87.74
12	87.32
16	85.31
20	82.21
24	81.34
28	80.32
32	78.61
36	75.43

The table 1 shows that the accuracy of Bi-LSTM model decreases with increment in topics. The topics 24 with accuracy 81.34% looks good for classification as the accuracy of model is quite good and also 24 topics shows clear insight into the topics of the data. The separation of topics is also good for 24 topics. Beyond 24 the topics start mixing up with each other.

The table 2 shows some important and most occurring words distinctly identifying each of 24 topics.

Table 2. Words identifying corresponding topics

Topics	**Words**
Topic 1	india, nagpur
Topic 2	style, life
Topic 3	hyderabad, india
Topic 4	visakhapatnam, gurgaon
Topic 5	world, cricket

Topic 6	india, ghaziabad
Topic 7	entertainment, movies
Topic 8	meerut, bareilly
Topic 9	crore, worth
Topic 10	citizen, stories
Topic 11	delhi, chennai
Topic 12	sports, india
Topic 13	business, india
Topic 14	woman, dies
Topic 15	news, tv
Topic 16	punjab, haryana
Topic 17	education, mumbai
Topic 18	jaipur, lucknow
Topic 19	bollywood, hindi
Topic 20	rain, kerala
Topic 21	patna, bihar
Topic 22	kolkata, india
Topic 23	ahmedabad, gujarat
Topic 24	dehradun, uttarakhand

4.2.4 Examples of some Sentiment and Topic prediction

a) The bomb on plane threat creates flutter at Bengaluru airport

- Sentiment prediction: Negative
- Topic prediction: Topic 10

b) India to see major growth in manufacturing sector in the year

- Sentiment prediction: Positive
- Topic prediction: Topic 3

5. Conclusion

The comparative analysis of supervised and unsupervised learnings was performed for text analysis. For supervised learning, sentiment analysis was performed on labeled data of news headlines and it was found that Bi-LSTM model outperformed the machine learning ensemble model with accuracy 84.92% against 81.67% accuracy of ensemble model.

Similarly, for unsupervised learning, two algorithms were used for topic modeling i.e., LSA and LDA and it was found that LDA performed more homogeneous and balanced distribution of topics in comparison to LSA. The topics were more distinctly visible in case of LDA than LSA as seen through clustering graphs for both the algorithms. Also, in case of LDA it was necessary to make a trade-off between accuracy and number of topics for better topic classification and considering this, 24 topics with accuracy 81.34% was found to be a better choice.

6. Acknowledgement

We would like to express our sincere gratitude to Dr. Nanda Bikram Adhikari for providing his invaluable guidance in completion of this research work.

References

[1] M.E. Sunil, S. Vinay, S, "Kannada Sentiment Analysis using vectorization and Machine Learning", Advances in Intelligent Systems and Computing, vol. 1408, 2021

[2] S.T. Rabani, Q.R. Khan, A.M.U.D. Khanday, "Detection of suicidal ideation on twitter using machine learning and ensemble approaches", Baghdad science journal, 17(4):1328-1339, 2020, doi: http://dx.doi.org/10.21123/bsj.2020.17.4.1328

[3] A.M. Schoene, G. Lacey, A.P. Turner, N. Dethlefs, "Dilated LSTM with attention for classification of suicide notes", Proceedings of the 10th international workshop on health text mining and information analysis, 136-145, 2019, doi: https://doi.org/10.18653/v1/D19-62

[4] A.C. Fernandes, R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, D. Chandran, "Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing", Scientific reports, 2018, doi: 10.1038/s41598-018-25773-

[5] M. Taboada, J. Brooke, M. Tofiloski, K. V. M. Stede," Lexicon-Based Methods for Sentiment Analysis", 1 Association for Computational Linguistics,2011

[6] J. Kamps, M. Marx, R.J. Mokken, M. de Rijke," Using WordNet to Measure Semantic Orientations of Adjectives", Language & Inference Technology Group, University of Amsterdam, 2001

[7] G. Xu, Y. Meng, X. Qiu, Z. Yu, X. Wu, "Sentiment analysis of comments text based on BiLSTM", IEEE access, vol. 7, pp. 51522-51532, 2019

[8] U. Chauhan, A. Shah, "Topic Modeling using Latent Dirichlet Allocation: A survey", ACM Computing surveys, vol. 54, issue 7, Sep, 2021

[9] H. Jelodar, Y. Wang, "Latent Dirichlet Allocation (LDA) and Topic Modeling: models, applications", Nov, 2017

[10] I. Vayansky, S.A.P. Kumar, "A review to topic modeling methods", Information Systems, vol. 94, Dec, 2020

Author's biography

Vijay Yadav currently pursuing his Masters in Computer System and Knowledge Engineering from Pulchowk Campus, Lalitpur, Nepal.

Subarna Shakya received MSc and PhD degrees in Computer Engineering from the Lviv Polytechnic National University, Ukraine, in 1996 and 2000, respectively. Currently, he is a professor at the Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University. His research interest includes e-government system, distributed and cloud computing, and software engineering and information system, Deep Learning, and Data Science.